information

Article

Utilizing Machine Learning for Context-Aware Digital Biomarker of Stress in Older Adults

Md Saif Hassan Onim [1], Himanshu Thapliyal [1,*] and Elizabeth K. Rhodus [2]

[1] VLSI Emerging Design and Nano Things Security Lab. (VEDANTS-Lab), Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA; monim@vols.utk.edu
[2] Sanders–Brown Center on Aging, Department of Behavioral Science, University of Kentucky, Lexington, KY 40536, USA; elizabeth.rhodus@uky.edu
* Correspondence: hthapliyal@utk.edu

Abstract: Identifying stress in older adults is a crucial field of research in health and well-being. This allows us to take timely preventive measures that can help save lives. That is why a nonobtrusive way of accurate and precise stress detection is necessary. Researchers have proposed many statistical measurements to associate stress with sensor readings from digital biomarkers. With the recent progress of Artificial Intelligence in the healthcare domain, the application of machine learning is showing promising results in stress detection. Still, the viability of machine learning for digital biomarkers of stress is under-explored. In this work, we first investigate the performance of a supervised machine learning algorithm (Random Forest) with manual feature engineering for stress detection with contextual information. The concentration of salivary cortisol was used as the golden standard here. Our framework categorizes stress into No Stress, Low Stress, and High Stress by analyzing digital biomarkers gathered from wearable sensors. We also provide a thorough knowledge of stress in older adults by combining physiological data obtained from wearable sensors with contextual clues from a stress protocol. Our context-aware machine learning model, using sensor fusion, achieved a macroaverage F-1 score of 0.937 and an accuracy of 92.48% in identifying three stress levels. We further extend our work to get rid of the burden of manual feature engineering. We explore Convolutional Neural Network (CNN)-based feature encoder and cortisol biomarkers to detect stress using contextual information. We provide an in-depth look at the CNN-based feature encoder, which effectively separates useful features from physiological inputs. Both of our proposed frameworks, i.e., Random Forest with engineered features and a Fully Connected Network with CNN-based features validate that the integration of digital biomarkers of stress can provide more insight into the stress response even without any self-reporting or caregiver labels. Our method with sensor fusion shows an accuracy and F-1 score of 83.7797% and 0.7552, respectively, without context and 96.7525% accuracy and 0.9745 F-1 score with context, which also constitutes a 4% increase in accuracy and a 0.04 increase in F-1 score from RF.

Keywords: CNN; machine learning; stress detection; context; cortisol; digital biomarkers

check for updates

Citation: Onim, M.S.H.; Thapliyal, H.; Rhodus, E.K. Utilizing Machine Learning for Context-Aware Digital Biomarker of Stress in Older Adults. *Information* **2024**, *15*, 274. https://doi.org/10.3390/info15050274

Academic Editors: Haifeng Wang, Norma B. Ojeda, Lu He and Andreas Triantafyllidis

Received: 16 February 2024
Revised: 2 April 2024
Accepted: 10 May 2024
Published: 12 May 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Stress monitoring plays a crucial role in healthcare, as it impacts individuals' cognitive functions and decision-making abilities [1]. It significantly influences work efficiency, and prolonged exposure to excessive stress and negative emotions can detrimentally affect both physical and mental well-being [2]. This is especially pertinent for individuals in specific professions and specific age groups. In addition to its direct impact on physical health, stress detection for older people is vital for maintaining their emotional and psychological well-being. The aging process often brings about significant life changes, such as loss of loved ones, retirement, or limitations in mobility, which can all contribute to feelings of stress, anxiety, and even depression. Left unaddressed, these emotional stressors can

significantly diminish an individual's overall quality of life and lead to a decline in social interactions and engagement with activities. During the pandemic alone, approximately 40% of adults in the United States have reported experiencing symptoms of stress, anxiety, or depressive disorder, and older adults are more susceptible to the impacts of stress compared to other age groups [3]. Figure 1 illustrates that the Alzheimer's Disease (AD) spectrum, encompassing a gradual progression from subtle cognitive decline through various stages to dementia, may be influenced by chronic elevations in cortisol concentration [4]. Stress-induced physiological responses, which are measurable as neurophysiological biomarkers, can offer valuable insights for understanding and potentially predicting AD progression. With reliable digital biomarkers, older adults can learn more about their stress patterns and become more adept at identifying triggers. Furthermore, having access to reliable biomarkers enables older persons to take charge of their stress management by embracing self-care routines, getting prompt medical attention, and establishing healthier lifestyle patterns. Additionally, making knowledgeable decisions as a consumer regarding healthcare products and services can improve results and raise satisfaction with the patient's overall experience. It not only enhances personal wellbeing but also empowers older populations by fostering a culture of health literacy and active participation in self-management of their own health and wellness.

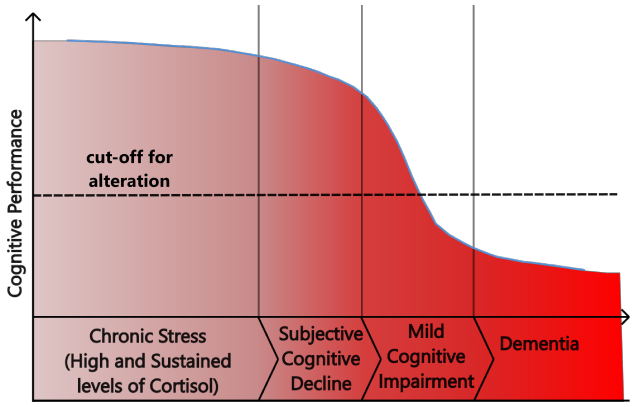

Figure 1. Cognitive performance vs. AD continuum due to chronic stress [4].

Digital biomarkers for stress refer to measurable and quantifiable indicators of stress levels that are collected through various sensors. These biomarkers provide insights into an individual's physiological or behavioral responses to stress, thus offering a more objective and real-time assessment compared to traditional self-reporting methods. Commonly employed digital biomarkers for detecting stress include Electrodermal Activity (EDA), Photoplethysmography (PPG), respiration, Electroencephalogram (EEG), and Skin Temperature (ST) [2]. Multiple studies are showing the differences in digital biomarker responses in different moods or stress levels [5–7].

Stress states can also manifest in behaviors such as eye movements, facial expressions, and sleeping and speech patterns [8]. These measures have been extensively utilized in laboratory settings for stress detection. Physiological signals are more practical to collect and offer greater objectivity than behavioral indicators. The advent of wearable sensors has provided a solution, thus enabling the use of wrist-worn devices for portable, inconspicuous, and noninvasive collection of physiological signals. Again, different combinations of signals from theses sensors also effect the outcome of stress response. Studies have shown the added benefit of using sensor fusion. Sensor fusion indicates the usage of combined features from multiple sensors that can dictate the output. Exercise, walking, and running, on the other hand, can generate a stress-like reactions, even though they are not detrimental to health. To address the problem, context-based machine learning algorithms that combine context with digital biomarkers are being developed to distinguish between different sources of stress [9]. Context typically refers to the information surrounding a particular event or situation that can be leveraged to make better predictions or decisions. This

surrounding information can include a wide range of factors depending on the application. Subjective stress is used as the ground truth, or gold standard, in these investigations, which are mostly directed toward younger individuals. On the other hand, according to studies, salivary cortisol is a prominent gold standard for stress biomarkers and should be used for ground truth estimation [10] instead of questionnaires or surveys filled out by third parties. But this approach is more suited for laboratory setting only and harder to obtain experimental results in real life settings.

In this work, we first test the performance of Random Forest, a machine-learning model for stress detection, with cortisol concentration as the ground truth. We also incorporate contextual information recorded during the test protocol. After the trial with manual feature engineering, we also automate the feature selection technique with a Convolutional Neural Network (CNN)-based feature encoder.

Here are the highlights of our contributions:

- We explore the effectiveness of machine learning models to find the correlation of digital biomarkers of stress with experimental data from 40 healthy older adults.
- We propose a ground truth labeling scheme based on cortisol concentration. We labeled stress into three distinct classes: No Stress, Low Stress, and High Stress.
- We investigate the efficacy of digital biomarkers from three signal streams (EDA, Blood Volume Pulse (BVP), and Interbeat Interval (IBI)) for stress classification.
- We also validate that the combinations of features from different sensors, also known as sensor fusion, enhances the accuracy of the machine learning classifier when compared to the case of a single-signal stream.
- We also propose a CNN-based feature encoder that automates the feature selection process and selects the best possible inputs for the FCN.
- We finally report that there's an increase in accuracy and F-1 score for the CNN-based feature extraction compared to the stress detection method with Random Forest on our dataset.

2. Related Works

In recent years, the advancement of machine learning has led to the gradual integration of deep networks into the realms of stress detection and emotion recognition based on digital biomarkers. Digital biomarkers not only enable accurate stress monitoring but also pave the way for personalized interventions. With continuous data collection, algorithms can learn individual patterns and preferences, thus tailoring interventions to suit each person's unique needs. This personalized approach has the potential to enhance the effectiveness of stress management strategies, thus moving away from generic solutions to more targeted and impactful interventions.

For instance, Saylam et al. [6] conducted a study on stress biomarkers aiming to predict stress levels by identifying crucial parameters from various modalities, including mobile phones and wearables. They utilized a ranking system to emphasize the significance of different modalities, such as sleep and activity levels, in classifying stress. Daily stress responses served as labels, thereby categorizing participants into five stress levels. The study compared the efficacy of considering only important parameters versus using all parameters by employing the Random Forest (RF) algorithm. Additionally, the researchers explored the potential impact of participants' personalities on stress levels to enhance overall understanding. Opoku et al. [5] conducted an analysis using a longitudinal dataset and employed statistical and machine learning methods to explore the connection between digital biomarkers, mood ratings, and depression. The study aimed to determine if there are differences in digital biomarkers and mood ratings between depressed and nondepressed participants. The findings revealed that although it was possible to accurately predict the depression status of participants with only digital biomarkers, a combination of digital biomarkers and mood ratings can increase the performance even more.

Neural Networks have also been employed in this field in recent times. By converting multimodal sequence signals into pictures and using an integrated, scalable, low-power

Deep Convolutional Neural Network (DCNN) to learn common features, Jafari et al. [11] achieved an impressive 94% accuracy rate in stress detection. In order to extract features from physiological data and behaviors gathered by wearable sensors, Aristizabal et al. [12] used deep networks. In order to extract deep features from EDA, PPG, and Zygomaticus Electromyography (zEMG), Hassan et al. [13] used the Deep Belief Network. These characteristics were then merged with statistical features. In order to extract their image features, Siddharth et al. [14] used deep networks to create spectrograms from every signal channel.

Despite the remarkable performance improvements achieved by deep learning-based algorithms in stress assessment, they come with only a few number of subjects and struggle to meet the requirements necessary for effectively training deep networks. Researchers have addressed such problems and proposed solutions based on machine learning with manual statistical feature engineering and have targeted specific age groups for less spread in the training set [15–17]. These studies have relied on questionnaires or annotations by external observers for ground truth on the base level and stress level. However, such ground truths are prone to error and differ from person to person. Hence, Nath et al. [18] used cortisol for the label and the EDA and PPG biomarker with context as an added input for stress classification. Table 1 shows an overview of recent works that deal with stress detection for older adults and takes context information into account.

Table 1. Summary of recent research on older adult stress detection.

Authors & Year	Signals	Feature Extraction	Base Model	Used Context	Stress Ground Truth
Ferreira et al. [19] 2014	ECG, EEG, EDA	Manual	Quadratic Discriminant Analysis	✗	Questionnaire
Kikhia et al. [20] 2016	EDA	Manual	Thresholding	✗	Annotation by clinical Staff
Belk et al. [15] 2016	EDA, IMU, HR, ST, Grip Force	Manual	Bayesian Probability	✗	Annotation by external observer
Delmastro et al. [16] 2020	ECG, EDA	Manual	Support Vector Machine,K-NN, Decision Tree	✗	Predetermined annotation
Cheong et al. [17] 2020	ST, HR, SC, HUM, AT	Manual	Statistical Correlation	✗	Questionnaire
Nath et al. [18] 2021	EDA, BVP, IBI, ST	Manual	Decision Tree	✗	Cortisol Concentration
Proposed Work 2024	EDA, BVP, IBI, ST	Manual	Random Forest	✓	Cortisol Concentration
		Automatic	1-D CNN	✓	

ECG: Electrocardiogram. EEG: Electroencephalography. EDA: Electrodermal Activity. SCL: Skin Conductance Level. IMU: Inertial Measurement Unit. HR: Heart Rate. ST: Skin Temperature. SC: Step Counter. HUM: Humidity. AT: Air Temperature, BVP: Blood Volume Pulse. IBI: Inter Beat Interval.

For quite some time, the scientific community has acknowledged the importance of monitoring stress levels based on physiological signals. The majority of currently released works, however, are focused on younger adults or fail to take advantage of relevant contexts. However, the literature in this domain shows that biomarkers like oxidative stress, cortisol, and metabolic imbalances are highly correlated with age and are fairly distinguishable among the age groups [21]. Similarly, the use of cortisol concentration as a stress biomarker has also been investigated. To calculate stress levels, the majority of studies have relied on subjective procedures or surveys. While there have been recent experiments in context-based stress detection, feature selection requires significant human involvement. The next step toward improvement will require an automatic feature encoder that employs context information and cortisol as the ground truth. With that in mind, we incorporated a 1D

CNN-based feature encoder. CNNs are a type of artificial neural network (ANN) with convolutional layers in their hidden layers [22]. They were first used in computer vision tasks, but they have since piqued the interest of researchers in a wide range of fields, including biosignal classification such as electromyogram (EMG) signal classification for gesture recognition and electroencephalogram (EEG) pattern identification for assistive machine control [23–26]. The CNN has been used effectively in arrhythmia detection [26], signal component identification [27], biometric recognition [28], and other applications in ECG signal classification. These experiments revealed the power of CNN in biomarker categorization, thus demonstrating that CNN has the capacity to identify stress. They also show the requirements for an for an end-to-end CNN-based context-aware machine learning model for stress biomarker in older persons. Therefore, in our work, we aim to experiment the aforementioned under-explored domains of stress biomarkers. We employ both feature engineered machine learning and CNN-based fully connected networks for stress biomarkers of older adults in laboratory settings and validate the stress labels with cortisol concentration.

3. Data Collection, Preprocessing, and Labeling

In this section, we will first explain the procedure of our data collection from the selected participants. Then, we mention the stress protocol that was incorporated to induce stress in participants. Then, we provide our ground truth labeling technique and its visualization after labeling. Afterwards, we explain the incorporation of contextual information followed by feature selection and preprocessing.

3.1. Data Collection

In our investigation, a group comprising 40 elderly individuals ranging in age from 60 to 80 years was selected as the subject population. Within this sample, the demographic distribution is comprised of 28 female and 12 male participants. As we are targeting a specific age group, 40 participants should be adequate for our experiments. Due to data corruption, the dataset of one participant was deemed unsuitable for inclusion, thus resulting in a final dataset derived from 39 participants. To provide an unbiased dataset, a thorough screening procedure was conducted on each participant prior to their registration in the experimental protocol to determine whether they had any pre-existing medical illnesses or disorders. We collected physiological data using the Empatica E4 wristband, which has built-in sensors for skin temperature (ST), photoplethysmography (PPG), and electrodermal activity (EDA), as illustrated in Figure 2.

Figure 2. Empatica wristband for data collection.

3.2. Stress Protocol: Trier Social Stress Test

As a stress inducer, the Trier Social Stress Test (TSST) protocol was employed for the experiment. The TSST, a widely recognized and established experimental framework, is distinguished for its capacity to stimulate stress within a context that emulates naturalistic conditions [29]. The procedural sequence of the TSST encompasses distinct phases, as shown in Figure 3. It consists of the waiting interval, prestress phase, stress induction period, and subsequent recovery phase. As part of the preparatory stage, participants engage

in the completion of demographic questionnaires. After the waiting period, the prestress (PS) phase starts, during which the base measurements are acquired. The collective duration of the waiting interval and the prestress phase comprise a duration of 20 min (T1–T2). Then, the prestress and stress phases span a duration of 20 min (T2–T3). The stress-inducing session is partitioned into three discrete segments: an initial 10 min span designated as the anticipatory stress phase (AS), which is succeeded by a 5 min interval encompassing speech and cognitive arithmetic exercises (M). During the anticipatory stress phase (AS), participants are tasked with a continuous discourse on a designated topic for 5 min while being observed. Then, they are engaged in the cognitive arithmetic exercise that requires resolving basic addition and subtraction equations. Importantly, the complexity of the cognitive tasks increases progressively with each accurate solution. The study ends with the inclusion of two sequential 20 min recovery intervals (T3–T5). Throughout the course of the experiment, salivary samples were collected on five distinct occasions corresponding to the time points denoted as T1, T2, T3, T4, and T5.

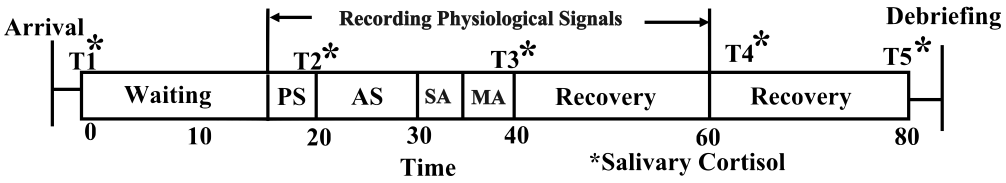

Figure 3. TSST protocol. PS: Prestress. AS: Anticipatory Stress. SA: Speech Assignment. MA: Mental Arithmetic. * refers to the time when saliva sample was collected.

3.3. Ground Truth Estimation from Cortisol Concentration

The participants' saliva samples were dispatched to the laboratory for cortisol concentration analysis. We classified cortisol levels into three distinct categories, namely No Stress, Low Stress, and High Stress, serving as the reference for addressing the multiclass classification problem. We designated the minimum value among the samples taken at time points T1, T2, and T5 as relaxed or No Stress, with the other two being categorized as Low Stress. The highest of the two samples collected at time points T3 and T4 was labeled as High Stress, while the remaining one was classified as Low Stress. Ultimately, these labels were expressed as numeric integers [30]. Figure 4 shows an example of cortisol concentration for a participant from the dataset. According to the algorithm mentioned before, the assigned labels would be No Stress: T-5, Low Stress: T-1, T-2, T-4, and High Stress: T-3. These reference labels were used as ground truth in laboratory settings for validation purposes.

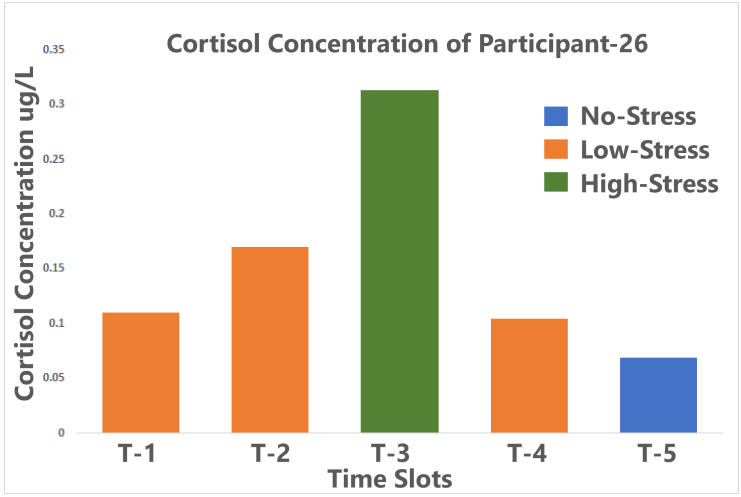

Figure 4. Cortisol concentration of a participant.

3.4. Distribution of Labeled Dataset and Incorporating Context

The labeled data were then projected with Fisher's Linear Discriminant (FLD) in Figure 5. It is a dimensionality reduction method that uses a projection vector $\vec{w}_p$ to reduce the dimension of the dataset from f to $n-1$, as shown in Equation (1).

$$\underset{M\times(n-1)}{X_{FLD}} = \underset{M\times f}{X^T_{original}} \cdot \underset{f\times(n-1)}{\vec{w}_p} \tag{1}$$

Here, M is the length of the data, f is the original number of features mentioned in the previous section, and n is the number of classes. It shows that our new ground truth labeling successfully differentiates the *No Stress* and *High Stress* classes, with some overlap between *Low Stress* and *High Stress*. There are also minor data points, which can be marked as outliers. These data points will be misclassified with any linear classifier unless any external information or context is provided to deal with them.

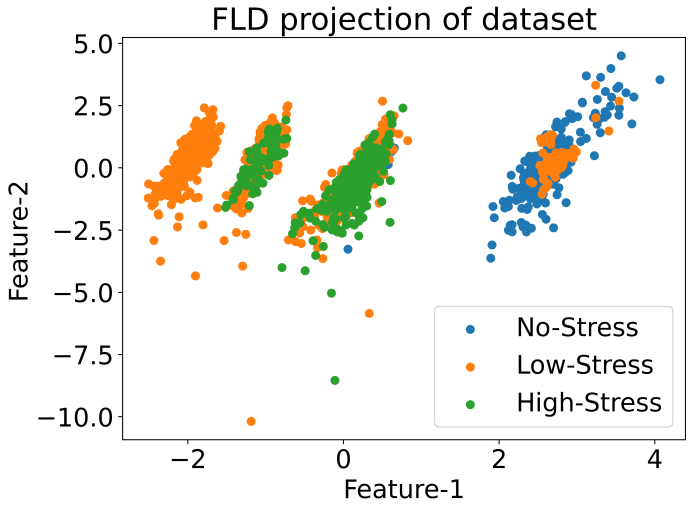

Figure 5. Distribution of projected 2D labeled dataset with FLD.

In this research, our contextual information was derived from the stress levels of participants before, during, and after the TSST. It is important to acknowledge that some participants may have arrived at the lab already experiencing stress levels higher than what was induced during the TSST, while others may have exhibited varying rates of stress level change following the TSST, thus leading to several challenges for our classifier. Primarily, these variations can result in an increased false positive rate, as the prior probability after ground truth estimation is not uniform across all classes, thus causing dataset imbalance. Additionally, the duration of the relaxed period may vary among individuals. To mitigate these unexpected outcomes and enhance decision making, we incorporated supplementary contextual information like the lab settings before, during, and after the TSST protocol directly into our dataset. These context features provided numeric values analogous to the ground truth labels, and their assignment was based on the laboratory conditions under which the data were collected.

4. Context-Aware Stress Detection with Supervised Machine Learning

Our method for context-aware stress detection with ML consists of three main parts. First, we have the statistical feature extraction, which we explain in Section 4.1. After collecting the data from the Empatica wristband, they go through preprocessing steps where they are filtered, normalized, and fused. This sensor fusion plays an important role in removing false positives, which we will show in more detail in the later part of the section. Second, the ground truth is estimated from the cortisol concentration, and the context

information comes from the TSST protocol. Finally, they are all fed to the Random Forest model for training. The overall workflow illustrated in Figure 6 is explained in this section.

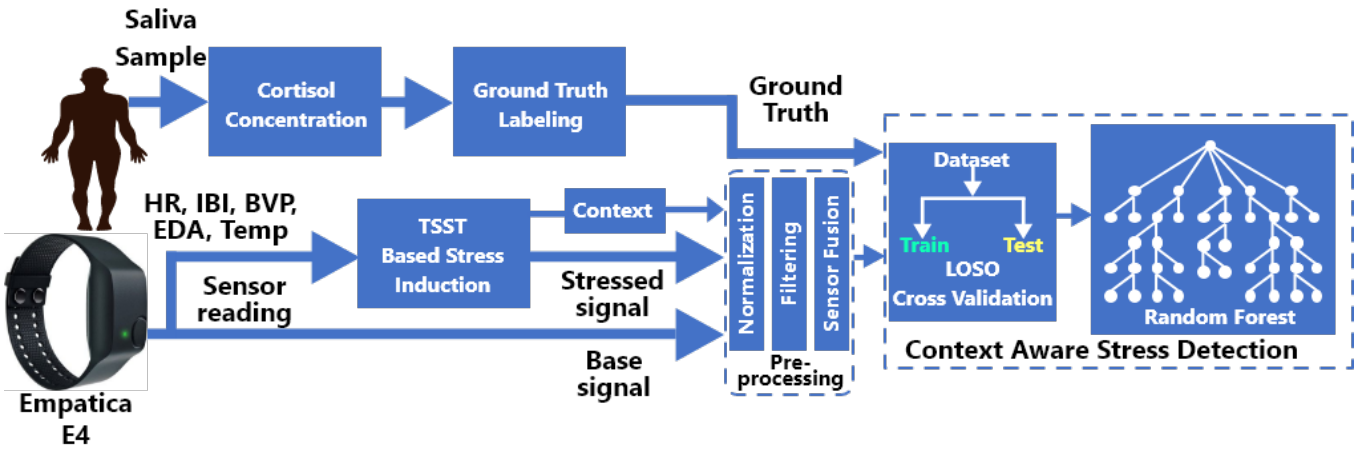

Figure 6. Workflow of stress detection with machine learning model (Random Forest).

4.1. Statistical Feature Extraction from Signal Stream

After collecting sensor data, they are preprocessed for the corresponding machine learning algorithm. The raw Electrodermal Activity (EDA) and Blood Volume Pressure (BVP) signals are normalized and filtered using a 5th-order low-pass filter with cut-off frequencies of 1 Hz for EDA and 10 Hz for BVP. The Interbeat Interval (IBI) signal undergoes no preprocessing. Subsequent to preprocessing, features are extracted using a window of 1.5 min, with an overlap of 0.75 min from the samples. Previous studies have indicated that characterizing EDA peaks or startles can provide insights into an individual's stress level [31]. The first derivative of the EDA signal undergoes a peak detection algorithm [32], and features such as peak amplitude, peak width, and peak prominence are computed. In total, 18 characteristics were extracted from the EDA signal. The BVP signal, obtained from the photoplethysmogram (PPG) sensor, reflects heart activity. The cardiovascular arousal associated with increased blood pressure and heart rate during stress is correlated with the BVP signal [33]. Seventeen characteristics were derived from the BVP signal, thus summarizing peak amplitude, breadth, and prominence similar to EDA features. Additionally, the IBI and ST signals contribute six time domain statistical characteristics each. The IBI, representing the interval between two heartbeats, is instrumental in understanding cardiac activity concerning stress. ST refers to the skin temperature in that instance.

4.2. Machine Learning Model

We have chosen Random Forest as our base model for the classification problem. Supervised machine learning algorithms, such as Random Forest, are commonly utilized for addressing classification and regression challenges. It constructs a multitude of decision trees during training and outputs the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. The algorithm is defined as follows:

1. Bootstrap Sampling:The dataset is broken into subsets to send randomly in individual trees.
2. Decision Trees: For each sample and feature subset, a decision tree is constructed. The splitting is based on a preset criterion. For the classification problem, we used logarithmic Gini impurity, which is shown in Equation (2):

$$Gini(D) = 1 - \sum_{i=1}^{C} (P_i)^2 \tag{2}$$

Here, D is the dataset at that tree node, C is the number of classes iterated by i, and P_i is the probability of class i in node D.

3. Ensembling and Voting: Based on the criterion, the required number of trees is constructed. The loss is calculated based on logarithmic entropy, as shown in Equation (3). These ensemble trees will provide a prediction of their own, and out of it, the majority-voted class will be the final prediction. The equation for the prediction of individual trees is shown in Equation (4), and the final prediction is shown in Equation (5). Here, $\hat{Pred}_i(x)$ denotes the prediction of a tree, $f(.)$ is the indicator function, and $L_i(x)$ is the leaf node to which x is assigned in tree i.

$$Entropy = -\sum_{i=1}^{C} P_i \times Log_2(P_i) \tag{3}$$

$$\hat{Pred}_i(x) = \text{argmax}_c \left(\sum_{x_j \in L_i(x)} f(y_j = c) \right) \tag{4}$$

$$Pred = argmax_i(\hat{Pred}_i) \tag{5}$$

4.3. Model Training and Testing

After extracting features from the data, we employed salivary cortisol analysis to label these features as indicative of three stress classes. To evaluate the performance, we utilized a Leave One Sample Out (LOSO) crossvalidation technique. This involved iteratively using one participant's data for testing while training on the remaining participants' data. During the training phase, a feature selection process based on Kendall's tau correlation was implemented to identify the most pertinent features. This correlation measure established the link between the training feature set and the stress labels derived from cortisol concentration [34]. Only features exhibiting statistically significant correlations (p values below 0.05) were retained. The final set of training features is presented in Table 2. When training the Random Forest model, hyperparameters were optimized based on the model's performance. The final selection of hyperparameters that resulted in the best accuracy on the test set is shown in Table 3.

Table 2. Selected Features from Physiological Signals.

Physiological Signals	Extracted Features	Statistical Measures
EDA	Amplitude:	Mean, Median, Maximum, Minimum Standard Deviation, Root Mean Square
	Width:	Median, Standard Deviation, Root Mean Square
	Prominence:	Minimum
BVP	Amplitude:	Mean, Standard Deviation, Root Mean Square
	Width:	Median
	Prominence:	Mean, Median, Maximum, Minimum, No of Peaks, Standard Deviation, Root Mean Square
IBI	Amplitude:	Maximum, Standard Deviation
ST	Amplitude:	Maximum, Standard Deviation

Table 3. Training parameters for Random Forest.

Parameter		Value
No. of Estimators	-	40
Criterion	-	Entropy
Minimum Samples Split	-	2
Maximum Depth	-	Till (minimum sample Split − 1)
Minimum Samples Leaf	-	1
Maximum Feature	-	Auto
Bootstrap	-	True
Random State	-	4

4.4. Result Analysis

We now assess the Random Forest model's performance using our dataset in terms of the estimated ground truth following training and testing. We will then talk about how well it performs with more *context*. We used the F-1 score and weighted accuracy to assess the model's performance.

The summarized results, as presented in Table 4, illustrate the performance of a single sensor and the sensor fusion. The enhancement in the F-1 score suggests that sensor fusion plays a role in mitigating the likelihood of false positive detections. Notably, our model achieved a commendable accuracy even in the absence of context, as per the proposed ground truth labeling. Introducing context resulted in an increase of around 20% in accuracy and around 0.2 in the F-1 score. The increase in the F-1 score explains the model's performance in not only making the correct detection for positive cases as well.

Table 4. Performance of Random Forest model with and without context.

Criteria	Sensor List	EDA	PPG	PPG	EDA, PPG	EDA, PPG	EDA, PPG, ST
	Signal List	**EDA**	**BVP**	**IBI**	**EDA, BVP**	**EDA, BVP, IBI**	**EDA, BVP, IBI, ST**
Features	**Total**	18	17	6	35	41	47
	Selected	11	11	2	22	24	27
Without Context	**Macro F-1 Score**	0.723	0.711	0.713	0.712	0.734	0.734
	Accuracy (%)	72.95	73.40	72.77	72.51	72.44	72.44
With Context	**Macro F-1 Score**	0.922	0.907	0.910	0.909	0.937	0.943
	Accuracy (%)	93.13	93.69	92.89	92.56	92.48	91.01

5. Context-Aware Stress Detection with CNN-Based Automatic Feature Encoder

Our method for context-aware stress detection with a CNN-based model is based on standard, well-known CNN architectures. First, we have the automatic feature encoder, which is a series of 1D convolutional layers. It is a sophisticated tool for extracting features and learning from 1D data sequences automatically. Based on loss, activation, and gradient values, we keep an appropriate number of neurons in each layer in our model architecture. During convolution, the CNN kernels move over the components of the 1D input signal stream. After collecting the data from the Empatica wristband, they are fed directly to CNN. Similar to ML, the ground truth is estimated from the cortisol concentration, and the context information comes from the TSST protocol. Finally, they are all fed to the feature encoder to extract a new feature representation. To test the performance for this case, we have used 5-fold crossvalidation. The overall workflow illustrated in Figure 7 is explained in this section.

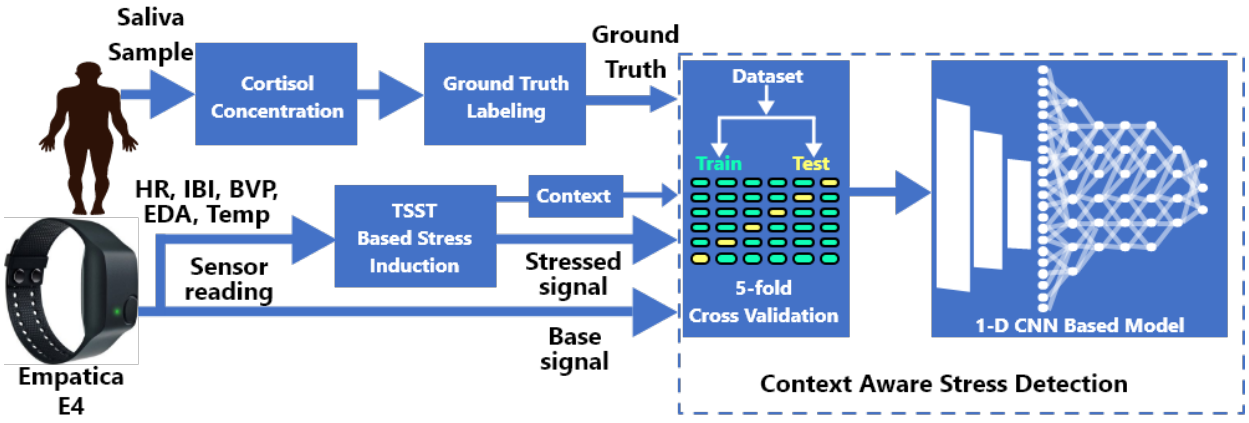

Figure 7. Workflow of stress detection with CNN-based model.

5.1. Automatic Feature Extraction with Feature Encoder

Our 1D CNN feature encoder consists of an *input* layer, a group of *convolutional* and *max pooling* layer pairs, a *flattening* layer, and *dense* layers. The *convolutional* layers have a gradually increasing number of filter sizes with a *Rectified Linear Unit* (ReLU) activation function followed by a *max pooling* layer. This type of pattern helps the model choose the best possible weights for the input streams. The ReLU activation function helps to avoid the vanishing gradient so that a faster convergence can be obtained, and the *max pooling* layer has been introduced to reduce the dimensions of the feature maps. Then, the *flattening* layer downsamples data into a 1D vector. Finally, there is a *dense–dropout* pair that reduces model overfitting. We have chosen the dropout rate to be 30% after trial and error for the best accuracy.

5.2. Fully Connected Neural Network

After the CNN-based feature encoder, the encoded features are sent to a series of *dense* and *dropout* layers for final decision making. The dropout rate was similar to the feature encoder with a ReLu activation. Finally, the output layer sets out three nodes with a softmax activation function that represents three stress classes.

As the problem formulated suggests it is a multimodal classification problem, Equation (8) provides the class prediction with output layer activation of *softmax* (σ_o), which is shown in Equation (6). Other layers have activation of *ReLu* (σ_h), as shown in Equation (7). The entropy, i.e., the loss to optimize, is chosen to be the *categorical crossentropy loss* shown in Equation (9).

$$\sigma_o(z) = \frac{e^{z_i}}{\sum_{j=1}^{N} e^{z_j}} \tag{6}$$

$$\sigma_h(z) = max(0, z) \tag{7}$$

$$class = arg\_max\left[\sigma\left\{(W_i)^T \times \phi(X_j) + b^i\right\}\right] \tag{8}$$

$$Loss(y, \hat{y}) = -\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) \tag{9}$$

Here, W_i, X_j, and b_i are the weight, inputs vector, and bias for ith and jth instances, respectively. N is the total number of samples, y_i is the true probability distribution (one-hot encoded) for class i, and $\hat{y}_i$ is the predicted probability for class i.

Our proposed CNN-based model consists of these two major parts cascaded together. Firstly, the feature encoder takes all possible statistical features and transforms them into a new feature representation. Secondly, these new features are fed to a fully connected neural network comprising several dense layers. The hyperparameters for this model is chosen similarly based on the best performing test dataset listed in Table 5. The primary determinant of a model's complexity is the quantity of trainable parameters. This comprises

the total number of neurons that are employed in creating the layer-to-layer connection. With m and n serving as input and output nodes, respectively, a dense layer with a total of $(n+1) \times m$ trainable parameters is formed. Convolutional layers with input and output feature maps of p and q, respectively, combined with a filter of size $i \times j$ result in total trainable parameters of $(i \times j \times p + 1) \times q$. We present a breakdown of the number of parameters in our suggested model in Table 6.

Table 5. Training parameters for convolutional neural network.

Parameter		Value
Base Architecture	-	CNN
Classes	-	3
Number of Epochs Trained	-	25
Hidden Layer Activation	-	Rectified Linear Unit (ReLU)
Output Layer Activation	-	Softmax
Optimizer	-	Adam
Loss Function	-	Categorical Crossentropy

Table 6. Proposed model layers.

Model	Layer Name	Layer Info	Number of Parameters
Feature Encoder	Conv-1D	Filter = 32	128
	Conv-1D	Filter = 64	6208
	Maxpool-1D	–	0
	Conv-1D	Filter = 128	24,704
	Maxpool-1D	–	0
	Conv-1D	Filter = 256	98,560
	Maxpool-1D	–	0
	Flatten	–	0
	Dense	Node = 2048	3,147,776
	Dropout	Rate = 30%	0
	Dense	Node = 512	1,049,088
	Dropout	Rate = 30%	0
Fully Connected NN	Dense	Node = 128	65,664
	Dense	Node = 32	4128
	Dense	Node = 8	264
	Dense	Node = 3	27
	Total Parameters		4,396,547

5.3. Result Analysis

Crossvalidation is a crucial technique in machine learning and statistical modeling that addresses the challenge of assessing a model's performance on new, unseen data. We performed 5-fold crossvalidation with an 80/20 train/test split and recorded the performance in each case. From Table 7, it can be seen that Fold-1 had the highest average accuracy, and Fold-3 had the highest macro F-1 score. On the other hand, Fold-2 had the lowest accuracy and F-1 score.

Again, for multiclass classification problems, the accuracy and F-1 score might not always provide a full picture of a model's performance. In such cases, the confusion matrix provides a quantitative measurement of the model's performance in each class. The confusion matrix generated from our experiment in Table 8 shows that only a few times did our model confuse Low Stress with High Stress. As seen from the projected dataset in Figure 5, this minor confusion was expected. The high value of its diagonal validates our model's performance on stress level classification.

Table 7. Performance for 5-fold crossvalidation.

Fold Sequence	Avg Accuracy	F-1 Score
Fold-1	99.8685% (highest)	0.9888
Fold-2	86.9696% (lowest)	0.9379 (lowest)
Fold-3	99.2796%	0.9928 (highest)
Fold-4	99.3216%	0.9752
Fold-5	98.3232%	0.9778
Average	96.7525%	0.9745

Table 8. Confusion matrix (weighted average of 5 folds).

Ground Truth	Predicted Class		
	No Stress	Low Stress	High Stress
No Stress	83.8	2.6	3
Low Stress	5.6	243.8	2.2
High Stress	2.6	2.8	58

Table 9 shows the performance of our proposed model with and without contextual information. It shows that the F-1 score increased from 0.7552 to 0.9745, and the accuracy increased from 83.7797% to 96.7525% after adding additional contextual information. The increased F-1 score suggests that sensor fusion and added context reduce the possibility of false positive detections. Based on the proposed neural network-based model, we obtained a reasonable accuracy without context, and after incorporating context, the accuracy increased by roughly 13%. With the sensor fusion listed in the last column, our model had an F-1 score of 0.75 and an accuracy of 83.78% when it performed without any contextual information from the TSST protocol. On the other hand, after the inclusion of contextual information, the F-1 increased to 0.97, which is 0.22 higher. Similarly, the accuracy increased by 13% to 96.75%.

Table 9. Performance of proposed CNN-based model with and without context.

Criteria	Sensor List	EDA	EDA, PPG	EDA, PPG	EDA, PPG, ST
	Signal List	EDA	EDA, BVP	EDA, BVP, IBI	EDA, BVP, IBI, ST
Without Context	Macro F-1 Score	0.6992	0.7161	0.7492	0.7552
	Accuracy (%)	80.0078	81.4290	83.2227	83.7797
With Context	Macro F-1 Score	0.9022	0.9240	0.9667	0.9745
	Accuracy (%)	92.3965	94.0378	96.1092	96.7525

6. Results Comparison and Discussion

Table 10 shows a comparison of the performance between the machine learning model with manual feature engineering and the CNN-based model with automatic feature selection. It is seen that the CNN-based model outperformed ML models in the case of both with and without contextual information. Although the F-1 score was similar, the accuracy for the CNN-based model was around 10% higher for cases without context and 4% higher for cases with context. The difference in performance mostly depends on the the type of data, best correlated features, and the architecture of the model itself. Random Forest worked better on tabular data, where it builds its own decision tress as weak learners to ensemble a final classifier. It can not modify the features on these parallel tress and thus highly depends on the input feature. Thus, the Random Forest model drew decision boundary for the given 27 features, which only resulted in some loss of information. On the other hand, CNN-based models are best suited for a larger number of data where in each layer, the features are given new representations based on the trainable parameters and

weights. By taking care of the overfitting problem, where the model tries to memorize the training set, neural networks can outperform many classical algorithms. For these reasons, the fully connected layers at the end have more correlated feature representation to make decisions and show increasing performance metrics.

Table 10. Performance comparison of signal fusion for proposed CNN-based model with machine learning model.

Criteria		Without Context		With Context	
Sensor List	**Signal List**	**Manual Feature ML**	**CNN-Based ML**	**Manual Feature ML**	**CNN-Based ML**
EDA, PPG, ST	EDA, BVP, IBI, ST	F-1: 0.73 ACC: 72.44	F-1: 0.75 ACC: 83.77	F-1: 0.94 ACC: 92.48	F-1: 0.97 ACC: 96.75

7. Conclusions and Future Works

Our findings and analyses shown in this paper demonstrate digital biomarkers as a viable option for developing a context-aware stress detection model for older persons. We have experimented with both feature-engineered and CNN-based machine learning models to find out their performance on stress biomarkers. The use of wrist-worn sensors to identify stress from digital biomarkers by correlating with cortisol can improve the state of the art. Our proposed algorithm can help bring clinical-level stress diagnosis into the ordinary consumer world, hence improving the quality of consumer health care. Our focus on a specific age group of older adults will help physicians prescribe personalized medications and treatments. This model also enables continuous stress monitoring, thus allowing an individual to regulate stress on their own. We provided two frameworks to validate our hypothesis on digital biomarkers. Our experimental results on the laboratory setting with cortisol concentration show the promise of utilizing machine learning for digital biomarkers of older adults. We also proved that the inclusion of context increased the machine learning model's performance metrics of stress detection for both of our proposed frameworks. Despite these developments, there is still space for improvement in the research that has been published thus far. Novel standardized datasets, crosscultural validation, the inclusion of more multimodal data sources, and real-time detection and predictions are a few of these requirements. In addition, for future directions, a careful analysis of the ethical implications of stress detection technology is necessary, including concerns about privacy and possible biases in algorithmic decision making.

Author Contributions: Conceptualization, M.S.H.O. and H.T.; methodology, M.S.H.O., H.T. and E.K.R.; software, M.S.H.O.; validation, M.S.H.O. and H.T.; formal analysis, M.S.H.O.; investigation, M.S.H.O., H.T. and E.K.R.; resources, H.T. and E.K.R.; data curation, M.S.H.O.; writing—original draft preparation, M.S.H.O. and H.T.; writing—review and editing, H.T. and E.K.R.; visualization, M.S.H.O. and H.T.; supervision, H.T.; project administration, H.T. and E.K.R.; funding acquisition, H.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Human Health and Wellness Research Development Program at the University of Tennessee, Knoxville, TN, USA.

Institutional Review Board Statement: All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Informed Consent Statement: Informed consent was obtained from all individual participants included in the study.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AD	Alzheimer's Disease
ANN	Artificial Neural Network
AT	Air Temperature
BVP	Blood Volume Pressure
CNN	Convolutional Neural Network
DCNN	Deep Convolutional Neural Network
ECG	Electrocardiogram
EDA	Electrodermal Activity
EEG	Electroencephalography
EMG	Electromyography
FLD	Fisher's Linear Discriminant
HR	Heart Rate
HUM	Humidity
IBI	Interbeat Interval
IMU	Inertial Measurement Unit
LOSO	Leave One Sample Out
PPG	Photoplethysmography
ReLU	Rectified Linear Unit
RF	Random Forest
SC	Step Counter
SCL	Skin Conductance Level
ST	Skin Temperature
TSST	Trier Social Stress Test
zEMG	Zygomaticus Electromyography

References

1. Harms, M.B. Stress and Exploitative Decision-Making. *J. Neurosci.* **2017**, *37*, 10035–10037. [CrossRef] [PubMed]
2. Giannakakis, G.; Grigoriadis, D.; Giannakaki, K.; Simantiraki, O.; Roniotis, A.; Tsiknakis, M. Review on Psychological Stress Detection Using Biosignals. *IEEE Trans. Affect. Comput.* **2022**, *13*, 440–460. [CrossRef]
3. Kourtis, L.C.; Regele, O.B.; Wright, J.M.; Jones, G.B. Digital biomarkers for Alzheimer's disease: The mobile/wearable devices opportunity. *NPJ Digit. Med.* **2019**, *2*, 9. [CrossRef]
4. Ávila Villanueva, M.; Gómez-Ramírez, J.; Maestú, F.; Venero, C.; Ávila, J.; Fernández-Blázquez, M.A. The Role of Chronic Stress as a Trigger for the Alzheimer Disease Continuum. *Front. Aging Neurosci.* **2020**, *12*, 561504. [CrossRef] [PubMed]
5. Opoku Asare, K.; Moshe, I.; Terhorst, Y.; Vega, J.; Hosio, S.; Baumeister, H.; Pulkki-Råback, L.; Ferreira, D. Mood ratings and digital biomarkers from smartphone and wearable data differentiates and predicts depression status: A longitudinal data analysis. *Pervasive Mob. Comput.* **2022**, *83*, 101621. [CrossRef]
6. Saylam, B.; Incel, O.D. Quantifying Digital Biomarkers for Well-Being: Stress, Anxiety, Positive and Negative Affect via Wearable Devices and Their Time-Based Predictions. *Sensors* **2023**, *23*, 8987. [CrossRef]
7. Jiang, Y.; Wang, W.; Scargill, T.; Rothman, M.; Dunn, J.; Gorlatova, M. Digital biomarkers reflect stress reduction after augmented reality guided meditation: A feasibility study. In *DigiBiom '22, Proceedings of the 2022 Workshop on Emerging Devices for Digital Biomarkers, Oregon, Portland, 1 July 2022*; Association for Computing Machinery: New York, NY, USA, 2022; pp. 29–34.
8. Giannakakis, G.; Pediaditis, M.; Manousos, D.; Kazantzaki, E.; Chiarugi, F.; Simos, P.; Marias, K.; Tsiknakis, M. Stress and anxiety detection using facial cues from videos. *Biomed. Signal Process. Control* **2017**, *31*, 89–101. [CrossRef]
9. Onim, M.S.H.; Rhodus, E.; Thapliyal, H. A Review of Context-Aware Machine Learning for Stress Detection. *IEEE Consum. Electron. Mag.* **2023**, 1–6. [CrossRef]
10. Payne, J.D.; Nadel, L. Sleep, dreams, and memory consolidation: The role of the stress hormone cortisol. *Learn. Mem.* **2004**, *11*, 671–678. [CrossRef]
11. Jafari, A.; Ganesan, A.; Thalisetty, C.S.K.; Sivasubramanian, V.; Oates, T.; Mohsenin, T. SensorNet: A Scalable and Low-Power Deep Convolutional Neural Network for Multimodal Data Classification. *IEEE Trans. Circuits Syst. I Regul. Pap.* **2019**, *66*, 274–287. [CrossRef]
12. Aristizabal, S.; Byun, K.; Wood, N.; Mullan, A.F.; Porter, P.M.; Campanella, C.; Jamrozik, A.; Nenadic, I.Z.; Bauer, B.A. The Feasibility of Wearable and Self-Report Stress Detection Measures in a Semi-Controlled Lab Environment. *IEEE Access* **2021**, *9*, 102053–102068. [CrossRef]
13. Hassan, M.M.; Alam, M.G.R.; Uddin, M.Z.; Huda, S.; Almogren, A.; Fortino, G. Human emotion recognition using deep belief network architecture. *Inf. Fusion* **2019**, *51*, 10–18. [CrossRef]

14. Jung, T.P.; Sejnowski, T.J. Utilizing Deep Learning Towards Multi-Modal Bio-Sensing and Vision-Based Affective Computing. *IEEE Trans. Affect. Comput.* **2022**, *13*, 96–107.
15. Belk, M.; Portugal, D.; Germanakos, P.; Quintas, J.; Christodoulou, E.; Samaras, G. A Computer Mouse for Stress Identification of Older Adults at Work. In Proceedings of the User Modeling, Adaptation, and Personalization, Halifax, NS, Canada, 13–17 July 2016.
16. Delmastro, F.; Di Martino, F.; Dolciotti, C. Cognitive Training and Stress Detection in MCI Frail Older People Through Wearable Sensors and Machine Learning. *IEEE Access* **2020**, *8*, 65573–65590. [CrossRef]
17. Cheong, S.M.; Bautista, C.; Ortiz, L. Sensing physiological change and mental stress in older adults from hot weather. *IEEE Access* **2020**, *8*, 70171–70181. [CrossRef]
18. Nath, R.K.; Thapliyal, H. Smart Wristband-Based Stress Detection Framework for Older Adults With Cortisol as Stress Biomarker. *IEEE Trans. Consum. Electron.* **2021**, *67*, 30–39. [CrossRef]
19. Ferreira, E.; Ferreira, D.; Kim, S.; Siirtola, P.; Röning, J.; Forlizzi, J.F.; Dey, A.K. Assessing real-time cognitive load based on psycho-physiological measures for younger and older adults. In Proceedings of the 2014 IEEE Symposium on Computational Intelligence, Cognitive Algorithms, Mind, and Brain (CCMB), Orlando, FL, USA, 9–12 December 2014; pp. 39–48.
20. Kikhia, B.; Stavropoulos, T.G.; Andreadis, S.; Karvonen, N.; Kompatsiaris, I.; Sävenstedt, S.; Pijl, M.; Melander, C. Utilizing a wristband sensor to measure the stress level for people with dementia. *Sensors* **2016**, *16*, 1989. [CrossRef]
21. Adeli, K.; Higgins, V.; Nieuwesteeg, M.; Raizman, J.E.; Chen, Y.; Wong, S.L.; Blais, D. Biochemical Marker Reference Values across Pediatric, Adult, and Geriatric Ages: Establishment of Robust Pediatric and Adult Reference Intervals on the Basis of the Canadian Health Measures Survey. *Clin. Chem.* **2015**, *61*, 1049–1062. [CrossRef]
22. Deng, L.; Yu, D. Deep Learning: Methods and Applications. *Found. Trends Signal Process.* **2014**, *7*, 197–387. [CrossRef]
23. Tabar, Y.R.; Halici, U. A novel deep learning approach for classification of EEG motor imagery signals. *J. Neural Eng.* **2016**, *14*, 016003. [CrossRef]
24. Zhai, X.; Jelfs, B.; Chan, R.H.M.; Tin, C. Self-Recalibrating Surface EMG Pattern Recognition for Neuroprosthesis Control Based on Convolutional Neural Network. *Front. Neurosci.* **2017**, *11*, 266372. [CrossRef]
25. Geng, W.; Du, Y.; Jin, W.; Wei, W.; Hu, Y.; Li, J. Gesture recognition by instantaneous surface EMG images. *Sci. Rep.* **2016**, *6*, 36571. [CrossRef]
26. Ruiz, J.T.; Pérez, J.D.B.; Blázquez, J.R.B. Arrhythmia Detection Using Convolutional Neural Models. In Proceedings of the Distributed Computing and Artificial Intelligence, 15th International Conference; Springer International Publishing: New York, NY, USA, 20–22 June 2018; pp. 120–127.
27. Xiang, Y.; Lin, Z.; Meng, J. Automatic QRS complex detection using two-level convolutional neural network. *Biomed. Eng. Online* **2018**, *17*, 13. [CrossRef]
28. Labati, R.D.; Muñoz, E.; Piuri, V.; Sassi, R.; Scotti, F. Deep-ECG: Convolutional Neural Networks for ECG biometric recognition. *Pattern Recognit. Lett.* **2019**, *126*, 78–85. [CrossRef]
29. Birkett, M.A. The Trier Social Stress Test Protocol for Inducing Psychological Stress. *J. Vis. Exp.* **2011**, *56*, e3238.
30. Onim, M.S.H.; Thapliyal, H. CASD-OA: Context-Aware Stress Detection for Older Adults with Machine Learning and Cortisol Biomarker. In *GLSVLSI '23, Proceedings of the Great Lakes Symposium on VLSI 2023, Knoxville, TN, USA, 5–7 June 2023*; Association for Computing Machinery: New York, NY, USA, 2023; pp. 103–108.
31. Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehlert, U. Discriminating stress from cognitive load using a wearable EDA device. *IEEE J. Biomed. Health Inform.* **2009**, *14*, 410–417. [CrossRef]
32. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
33. Kulkarni, S.; O'Farrell, I.; Erasi, M.; Kochar, M. Stress and hypertension. *WMJ Off. Publ. State Med Soc. Wis.* **1998**, *97*, 34.
34. Kendall, M.G. A new measure of rank correlation. *Biometrika* **1938**, *30*, 81–93. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.